# Some Accelerated Methods for Smooth Convex Minimization

**Alex L. Wang**, Daniels School of Business, Purdue University

Based on joint work with: **Benjamin Grimmer**, **Kevin Shu**

Mitchell E. Daniels, Jr.
School of Business

## Game plan/outline

- Problem setup:
  - smooth convex optimization
  - Fixed-Step First-Order Methods (FSFOMs)
- General fact about FSFOMs: existence of "shadow" iterate
- Recover Nesterov's **Fast Gradient Method** (FGM) and the **Optimized Gradient Method** (OGM) as "greedy" algorithms
- Develop **Subgame Perfect Gradient Method** (SPGM)

# Setup

# Smooth Convex Minimization

- Want algorithms for

$$\min_{x\in\mathbb{R}^d} f(x)$$

  - $f$ is convex
  - $f$ is 1-smooth: $\|\nabla f(x) - \nabla f(y)\| \leq \|x - y\| \quad \forall x, y$
  - $f$ has a minimizer $x_\star$ with minimum value $f_\star$
- $\mathcal{F}$ is the set of instances
- Learn information about $f \in \mathcal{F}$ via first-order queries

$$x \mapsto (f(x),\, \nabla f(x))$$

# Fixed-step first-order methods (FSFOM)

- $N$-step fixed-step first-order method (FSFOM)
- Strictly lower triangular matrix $\mathbf{H} \in \mathbb{R}^{[0,N]\times[0,N]}$

**Algorithm.** FSFOM($\mathbf{H}$)

- Initialize $x_0 = 0$
- For $n = 1, \dots, N$, iterate

$$x_n := x_{n-1} - \sum_{i=0}^{n-1} \mathbf{H}_{n,i} \nabla f(x_i)$$

- Output $x_N$

- Abbreviate $f_n = f(x_n)$ and $g_n = \nabla f(x_n)$
- Throughout talk, assume $d \geq N + 2$

# Performance

- Worst-case performance of $\mathbf{H}$:
  Define $r(\mathbf{H})$ to be largest value of $r$ s.t.

$$f_N - f_\star \leq \frac{1}{2r} \left\| x_0 - x_\star \right\|^2 \qquad \forall f \in \mathcal{F}$$

- Equivalently $$r(\mathbf{H}) := \min_{f \in \mathcal{F}} \frac{\frac{1}{2} \left\| x_0 - x_\star \right\|^2}{f_N - f_\star}$$

- The algorithm design problem:

$$\max_{\mathbf{H}} r(\mathbf{H})$$

- $N$ steps of gradient descent, $x_n = x_{n-1} - g_{n-1}$:

$$\mathbf{H} = \begin{bmatrix} 0 & & & & \\ 1 & 0 & & & \\ & 1 & 0 & & \\ & & \ddots & \ddots & \\ & & & 1 & 0 \end{bmatrix} \quad \text{and} \quad r(\mathbf{H}) = 2N+1$$

- [Nesterov 05]: $N$ steps of Fast Gradient Method (FGM):

$$\mathbf{H} = \begin{bmatrix} 0 & & & & \\ 1 & 0 & & & \\ 0 & 1.28175 & 0 & & \\ 0 & 0.122293 & 1.43404 & 0 & \\ 0 & 0.0649454 & 0.230504 & 1.53106 & 0 \end{bmatrix} \quad \text{and} \quad r(\mathbf{H}) \approx \frac{N^2}{4}.$$

# The “Shadow Iterate” and Acceleration

**Theorem.** [Grimmer Shu Wang 2024c]

Let $\mathbf{H}$ be any* $N$-step FSFOM. Let $r = r(\mathbf{H})$ so that

$$f_N - f_\star \leq \frac{1}{2r}\|x_0 - x_\star\|^2 \quad \forall f \in \mathcal{F}$$

We can construct a vector $v \in \mathbb{R}^{[0,N]}$ so that

$$f_N - f_\star + \frac{1}{2r}\left\|x_0 - \sum_{i=0}^{N} v_i g_i - x_\star\right\|^2 \leq \frac{1}{2r}\|x_0 - x_\star\|^2 \quad \forall f \in \mathcal{F}$$

- $z_{N+1} := x_0 - \sum_{i=0}^N v_i g_i$ is the shadow iterate for $\mathbf{H}$
- For any $\mathbf{H}$, $f \in \mathcal{F}$, either

$f(x_N) - f_\star$ outperforms worst-case or $z_{n+1} \approx x_\star$

# Example

- Let $N = 0$ and $\mathbf{H} = [0]$
- This “algorithm” outputs $x_0 = 0$ on any $f \in \mathcal{F}$
- By $1$-smoothness

$$f_0 - f_\star \leq \frac{1}{2} \|x_0 - x_\star\|^2$$

- Setting $v = [1]$

$$f_0 - f_\star + \frac{1}{2} \|x_0 - g_0 - x_\star\|^2 \leq \frac{1}{2} \|x_0 - x_\star\|^2$$

# Recovering Nesterov's FGM I

- **Idea**: Let's inductively derive a good FSFOM
- At iteration $n$, have $x_{n-1}$, $r_{n-1}$

$$f_{n-1} - f_\star \leq \frac{1}{2r_{n-1}} \left\| x_0 - x_\star \right\|^2$$

- For free, also have $z_n$ so that

$$f_{n-1} - f_\star + \frac{1}{2r_{n-1}} \left\| z_n - x_\star \right\|^2 \leq \frac{1}{2r_{n-1}} \left\| x_0 - x_\star \right\|^2$$

- Hedge between: either $f_{n-1} - f_\star$ already small or $z_n \approx x_\star$
- Let $\alpha_n \in (0,1)$ and set

$$x_n = \alpha_n \left( x_{n-1} - g_{n-1} \right) + (1 - \alpha_n) z_n$$

# Recovering Nesterov's FGM II

- **Goal**: Pick $(r_n, \alpha_n)$ so that

$$f(x_n) - f_\star \leq \frac{1}{2r_n} \|x_0 - x_\star\|^2$$

- **Ingredients**: $r_{n-1}, x_{n-1}, z_n$, and $f \in \mathcal{F}, x_\star$ satisfy

$$f(x_{n-1}) - f_\star + \frac{1}{2r_{n-1}} \|z_n - x_\star\|^2 \leq \frac{1}{2r_{n-1}} \|x_0 - x_\star\|^2$$

- **FGM**: Pick $\alpha_n$ to maximize worst-case $r_n$ for all $f \in \mathcal{F}$, $x_\star, x_{n-1}, z_n$ satisfying inductive hypothesis
- Explicit formula for $(r_n, \alpha_n)$ in terms of $r_{n-1}$

# Recovering Nesterov's FGM III

**Algorithm.** Fast Gradient Method [Nesterov 05]

- Initialize $x_0 = 0$, $z_1 = x_0 - g_0$, $r_0 = 1$

$$f_0 - f_\star + \frac{1}{2r_0}\left\|z_1 - x_\star\right\|^2 \leq \frac{1}{2r_0}\left\|x_0 - x_\star\right\|^2$$

- For $n = 1, \ldots, N$, set

$$\begin{aligned} x_n &= \alpha_n(x_{n-1} - g_{n-1}) + (1-\alpha_n)z_n \\ z_{n+1} &= \text{inductively maintained} \end{aligned}$$

where $\alpha_n$ greedily maximizes $r_n$ in

$$f_n - f_\star + \frac{1}{2r_n}\left\|z_{n+1} - x_\star\right\|^2 \leq \frac{1}{2r_n}\left\|x_0 - x_\star\right\|^2$$

- Output $x_N$ with performance $r_N^{\text{FGM}} \approx \frac{N^2}{4}$

# Recovering OGM I

- In FGM, each step starts with hypothesis
$$f_{n-1} - f_\star \text{ is small} \quad \text{or} \quad \|z_n - x_\star\|^2 \text{ is small},$$
but proof *actually* uses the fact that
$$f_{n-1} - \boxed{\frac{1}{2}\|g_{n-1}\|^2} - f_\star \text{ is small} \quad \text{or} \quad \|z_n - x_\star\|^2 \text{ is small}$$
- **Optimized Gradient Method** (OGM):
Suppose $x_{n-1}, r_{n-1}, z_n$ satisfy
$$f_{n-1} - \boxed{\frac{1}{2}\|g_{n-1}\|^2} - f_\star + \frac{1}{r_{n-1}}\|z_n - x_\star\|^2 \leq \frac{1}{r_{n-1}}\|x_0 - x_\star\|^2$$
- Pick $x_n = \alpha_n(x_{n-1} - g_{n-1}) + (1 - \alpha_n)z_n$

[Drori Teboulle 12] [Kim Fessler 16]

# Recovering OGM II

- **Goal**: Pick $(r_n, \alpha_n)$ so that

$$f(x_n) - \frac{1}{2}\left\|\nabla f(x_n)\right\|^2 - f_\star \leq \frac{1}{2r_n}\left\|x_0 - x_\star\right\|^2$$

- **Ingredients**: $r_{n-1}, x_{n-1}, z_n$, and $f \in \mathcal{F}, x_\star$ satisfy

$$\begin{aligned} f(x_{n-1}) - \frac{1}{2}\left\|\nabla f(x_{n-1})\right\|^2 - f_\star + \frac{1}{2r_{n-1}}\left\|z_n - x_\star\right\|^2 \\ \leq \frac{1}{2r_{n-1}}\left\|x_0 - x_\star\right\|^2 \end{aligned}$$

- **OGM**: Pick $\alpha_n$ to maximize worst-case $r_n$ for all $f \in \mathcal{F}$, $x_\star, x_{n-1}, z_n$ satisfying inductive hypothesis
- Explicit formula for $(r_n, \alpha_n)$ in terms of $r_{n-1}$

# Recovering OGM III

**Algorithm.** Optimized Gradient Method

- Initialize $x_0 = 0$, $z_1 = x_0 - 2g_0$, $r_0 = 2$

$$f_0 - \frac{1}{2}\|g_0\|^2 - f_\star + \frac{1}{2r_0}\|z_1 - x_\star\|^2 \leq \frac{1}{2r_0}\|x_0 - x_\star\|^2$$

- For $n = 1, \ldots, N-1$, set

$$x_n = \alpha_n(x_{n-1} - g_{n-1}) + (1 - \alpha_n)z_n$$

$$z_{n+1} = \text{inductively maintained}$$

where $\alpha_n$ greedily maximizes $r_n$ in

$$f_n - \frac{1}{2}\|g_n\|^2 - f_\star + \frac{1}{2r_n}\|z_{n+1} - x_\star\|^2 \leq \frac{1}{2r_n}\|x_n - x_\star\|^2$$

- Slight modification for iteration $N$ ...
- Output $x_N$ with performance $r_N^{\text{OGM}} \approx \frac{N^2}{2}$

# Optimality of OGM

**Theorem.** [Drori 2017]

The $N$-step Optimized Gradient Method (OGM) has rate

$$r_N^{\text{OGM}} \approx \frac{N^2}{2} \approx 2r_N^{\text{FGM}}$$

Furthermore, $N$-step OGM solves

$$\max_{\mathbf{H}} r(\mathbf{H}) = \max_{\mathbf{H}} \min_{f \in \mathcal{F}} \frac{\frac{1}{2}\left\|x_0 - x_\star\right\|^2}{f_N - f_\star}$$

# Subgame Perfect Gradient Method

## Can we do better than OGM?

- OGM is optimal in a uniform sense

$$f_N - f_\star \leq \frac{1}{2 \cdot r_N^{\text{OGM}}} \left\| x_0 - x_\star \right\|^2 \qquad \forall f \in \mathcal{F}$$

- Can perform at worst-case rate even on "easy instances"
- **Example**: $f(x) = \frac{1}{2} \left\| x - x_\star \right\|^2$ is a worst-case function!
  "Correct behavior" should terminate after two steps
- **Goal**: "Optimally tighten" performance of OGM

# The Convex Minimization Game I

- Model convex min. as sequential zero-sum game:
- Rounds $= N$, Alice $=$ Algorithm, Bob $=$ "adversary"
  - Alice plays $x_0 = 0$ and Bob plays $(f_0,\ g_0)$
  - At round $n = 1, \dots, N$

    Alice plays $x_n \in x_0 + \mathrm{span}(\{g_0, \dots, g_{n-1}\})$

    Bob plays $(f_n, g_n)$
  - Bob plays $(x_\star,\ f_\star,\ g_\star = 0)$, $f \in \mathcal{F}$ agreeing with history
- Alice's payoff is $$\frac{\frac{1}{2}\left\|x_0 - x_\star\right\|^2}{f_N - f_\star}$$

# The Convex Minimization Game II

- The OGM strategy is a Nash Equilibrium strategy
  - Alice's payoff $\geq r_N^{\text{OGM}}$
  - If Bob plays optimally, then no strategy for Alice can guarantee a payoff $> r_N^{\text{OGM}}$
- Subgame perfect notion captures idea of "optimally exploiting suboptimal play by adversary"
- OGM is not a subgame perfect Nash Equilibrium strategy
  - Consider $f(x) = \frac{1}{2}\left\|x - x_\star\right\|^2$
    Alice increases payoff $r_N^{\text{OGM}} \to \infty$ by deviating at iteration 2
- The "dynamic" extension of OGM is Subgame Perfect!

# SPGM Update

- **OGM**: Given $r_{n-1}$, set
  $x_n = \alpha_n(x_{n-1} - g_{n-1}) + (1-\alpha_n)z_n$ to maximize $r_n$ s.t.

  $\forall f \in \mathcal{F}, x_\star, x_{n-1}, z_n :$

$$f_{n-1} - \frac{1}{2}\|g_{n-1}\|^2 - f_\star + \frac{1}{2r_{n-1}}\|z_n - x_\star\|^2 \le \frac{1}{2r_{n-1}}\|x_0 - x_\star\|^2$$

$$\implies f_n - \frac{1}{2}\|g_n\|^2 - f_\star \le \frac{1}{2r_n}\|x_0 - x_\star\|^2$$

- **SPGM**: Given FO-history, set
  $x_n \in x_0 + \mathrm{span}(\{g_0, g_1, \ldots, g_{n-1}\})$ to maximize $r_n$ s.t.

  $\forall f \in \mathcal{F}, x_\star :$

$$f(x_i) = f_i, \quad \nabla f(x_i) = g_i, \qquad \forall i \in [0, n-1]$$

$$\implies f_n - \frac{1}{2}\|g_n\|^2 - f_\star \le \frac{1}{2r_n}\|x_0 - x_\star\|^2$$

  Can be reparameterized as a convex problem!

**Algorithm.** SPGM [Grimmer Shu Wang 2024b]

- Initialize $x_0 = 0$, $z_1 = x_0 - 2g_0$, $r_0 = 2$

$$f_0 - \frac{1}{2}\|g_0\|^2 - f_\star + \frac{1}{2r_0}\|z_1 - x_\star\|^2 \leq \frac{1}{2r_0}\|x_0 - x_\star\|^2$$

- For $n = 1, \ldots, N-1$, set

  $x_n, r_n, z_{n+1}$ = "output" of some convex minimization problem

  where $x_n, z_{n+1}$ greedily maximizes $r_n$ in:

$$f_n - \frac{1}{2}\|g_n\|^2 - f_\star + \frac{1}{2r_n}\|z_{n+1} - x_\star\|^2 \leq \frac{1}{2r_n}\|x_n - x_\star\|^2$$

- Slight modification for iteration $N$ ...
- Output $x_N$

**Theorem.** [Grimmer Shu Wang 24b]

**The SPGM is subgame perfect:**
Suppose Alice plays according to SPGM. After iteration $n$, Alice can guarantee a payoff of at least

$$r_N(\{(x_0, f_0, g_0), (x_1, f_1, g_1), \ldots, (x_n, f_n, g_n)\}) \geq r_N^{\text{OGM}}.$$

If Bob plays optimally in this subgame, then no strategy for Alice can guarantee a strictly larger payoff.

# Limited-memory SPGM

- SPGM overhead:
  - storage: $\{(x_n, f_n, g_n)\}$
  - solve convex problem in $2n$ variables optimally
- Limited-memory variant $k$-SPGM:
  - store $k$ tuples $\{(x_n, f_n, g_n, r_n, z_{n+1})\}$
  - solve convex problem in $2k$ variables
    - If optimal, then this is subgame perfect for the limited memory version of the minimization game
    - Correctness depends only on feasibility (there is a known feasible point corresponding to the OGM update)

# Summary/pointers

## Summary/pointers

- Existence of “Shadow Iterates” **See**: Blog post on my website
- **Recovered**: **See**: **[d’Aspremont Scieur Taylor 21]**
  - Nesterov’s Fast Gradient Method **See**: **[Nesterov 05]**
  - Optimized Gradient Method **See**: **[Drori Teboulle 12] [Kim Fessler 16] [Drori 17]**
- **New**: Subgame Perfect Gradient Method and limited-memory $k$-SPGM
  **See**: **[Grimmer Shu Wang 2024b]**
- Other recent work:
  - Acceleration without momentum “silver stepsize schedule”
    (possible if shadow iterate $z_{n+1} \in x_n + \mathbb{R} g_n$)
    **See**: **[Altschuler Parrilo 2023a,b] [Grimmer Shu Wang 2024a,c] [Zhang Jiang 2024]**

# References I


Altschuler, J. M. and Parrilo, P. A. (2023a). Acceleration by stepsize hedging i: Multi-step descent and the silver stepsize schedule.

Altschuler, J. M. and Parrilo, P. A. (2023b). Acceleration by stepsize hedging ii: Silver stepsize schedule for smooth convex optimization.

Drori, Y. (2017). The exact information-based complexity of smooth convex minimization. *Journal of Complexity*, 39:1–16.

Drori, Y. and Teboulle, M. (2012). Performance of first-order methods for smooth convex minimization: a novel approach. *Mathematical Programming*, 145:451–482.

d'Aspremont, A., Scieur, D., Taylor, A., et al. (2021). Acceleration methods. *Foundations and Trends® in Optimization*, 5(1-2):1–245.

Grimmer, B., Shu, K., and Wang, A. L. (2024a). Accelerated objective gap and gradient norm convergence for gradient descent via long steps.

Grimmer, B., Shu, K., and Wang, A. L. (2024b). Beyond minimax optimality: A subgame perfect gradient method. *arXiv preprint arXiv:2412.06731*.

Grimmer, B., Shu, K., and Wang, A. L. (2024c). Composing optimized stepsize schedules for gradient descent. *arXiv preprint arXiv:2410.16249*.

# References II

Kim, D. and Fessler, J. A. (2016). Optimized first-order methods for smooth convex minimization. *Math. Program.*, 159(1–2):81–107.

Nesterov, Y. (2005). Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1):127–152.

Zhang, Z. and Jiang, R. (2024). Accelerated gradient descent by concatenation of stepsize schedules.